ODYSSEY04 - The Speaker and
Language Recognition Workshop
Toledo, Spain
May 31 - June 3, 2004

Analysis of Multitarget Detection for Speaker and Language Recognition*

Elliot Singer and Douglas Reynolds

MIT Lincoln Laboratory
Lexington, MA, USA
{dar,es}@ll.mit.edu

Abstract

The general multitarget detection (open-set identification) task is the intersection of the more familiar tasks of close-set identification and open-set verification/detection. In the multitarget detection task, an input of unknown class is processed by a bank of parallel detectors and a decision is required as to whether the input is from among the target classes and, if so, which one. In this paper, we show analytically how the performance of a multitarget detector can be predicted from the open-set detection performance of the individual detectors of which it is constructed. We use this analytical framework to establish the relationship between the multitarget detector's closed-set identification error rate and its open-set detector miss and false alarm probabilities. Experiments performed using standard speaker and language corpora are described that demonstrate the validity of the analysis.

1. Introduction

The tasks of speaker and language recognition encompass both closed-set identification and open-set verification or detection. In closed-set identification, the aim is to associate an input utterance from an unknown speaker or language (or class in general) with one of a known set of speaker or language models. The task is referred to as closed-set since it assumed that the input utterance must be associated with a speaker or language known to the system a priori (i.e., one of the model classes). While the closed-set identification task is of academic interest and has potential diagnostic utility, there are few applications where one can legitimately assume a closed-set situation. Open-set detection, on the other hand, aims to determine or detect if an input utterance is from a specified speaker a specified language. The task is open-set since the input test utterance can be from a class unknown to the system a priori.

The most general recognition case is the intersection of closed-set identification and open-set detection and is known as open-set identification or multitarget detection. In multitarget detection, two questions must be answered: first, does the input utterance belong to one of the target classes and, if so, which class does it come from; i.e., the task consists of detection followed by identification. The multitarget detection task has not been the subject of as much study in the speech literature as the pure identification and detection/verification tasks have, although there has been some work in this area in the biometrics and face recognition fields [1,2]. Multitarget detection has application to the audio database search task for recorded meetings, broadcast news, or historical audio documents, where one may wish to make queries for a set of speakers or languages.

In this paper we will analyze theoretically and empirically the behavior of multitarget detectors using speaker and language data and tasks for experiments. The aim of this study is to characterize the performance of a multitarget detector on the basis of the performance of the individual target detectors of which it is constructed, and we will use this framework to derive the relation between the closed-set identification error and the open-set detector miss and false alarm errors.

The next section provides the framework and definitions of the multitarget detector. Section 3 derives equations for predicting the performance of a multitarget detector from the performance of single target detectors. In Section 4 we present experimental results for speaker and language data and tasks that show how well empirical performance matches predicted performance. Conclusions and future directions are provided in the final section.

2. Description of multitarget detector

2.1. Single target detector

We approach the analysis of multitarget detection by treating it as a generalization of the familiar single target class detection system shown in Figure 1, where a detector of class C processes the input test utterance x of class C_x and produces a similarity score y. An accept/reject decision is made by comparing the score y to a decision threshold θ. A decision to accept the input ($y>\theta$) is accompanied by the hypothesis h that the input is of class C. A miss error occurs on a target trial if the score y is below threshold

$$Miss:\; y<\theta \mid C_x = C \qquad (1)$$

and a false alarm error occurs on a non-target trial if the score is above threshold

$$FA:\; y>\theta \mid C_x \neq C \qquad (2)$$

Consequently, the single detector miss and false alarm probabilities are given by

$$P_{miss}(\theta)=\Pr[y<\theta \mid C_x = C] \qquad (3)$$

and

$$P_{fa}(\theta)=\Pr[y>\theta \mid C_x \neq C] \qquad (4)$$

In the speech community, detector performance is typically reported using a detection error tradeoff (DET) plot [3] that specifies the measured false alarm and miss rate values at all score thresholds.

* This work is sponsored by the Department of Defense under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

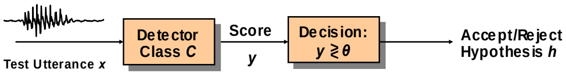

Figure 1: Single detector block diagram.

2.2. Multitarget (stack) detector

A block diagram of a multitarget (or stack) detector is shown in Figure 2. We assume the existence of a stack of S single class detectors for classes $C_1,\ldots,C_S$ operating in parallel on an input test utterance x of unknown class C_x. The set of classes represented by the detectors is also referred to as a watch list in the literature [4]. The detectors produce a set of scores $y_1,\ldots,y_S$ and a corresponding set of class hypotheses $h_1,\ldots,h_S$. The S scores are ranked and a verification decision is made by comparing the maximum score y^* to a decision threshold θ. A decision to accept leads to acceptance of the top k hypotheses $\{h_k^*\}$ (i.e., the hypotheses that correspond to the k highest scores), where $1 \le k \le S$. A decision to reject leads to rejection of all k hypotheses.

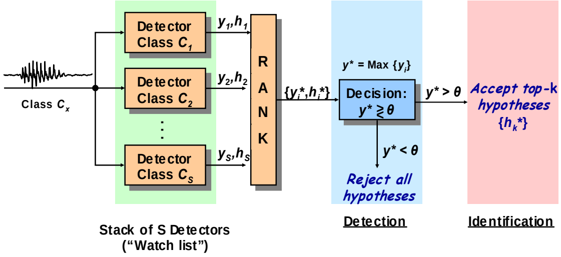

Figure 2: Multitarget (stack) detector block diagram.

2.3. Multitarget detector errors

By analogy to the definition of a false alarm for a single detector, we define a false alarm FA' of a multitarget detector as occurring when the maximum score y^* is above threshold given that the test message x is of an imposter (non-target) class

$$FA':\ y^* > \theta \mid C_x \notin \{C_{1,\ldots,S}\} \tag{5}$$

Thus, for imposter inputs and a given threshold θ, a false alarm occurs when one or more of the individual detectors produce a false alarm. Note that the occurrence of a false alarm is independent of the choice of k.

A miss can also be treated by extension of the single detector case

$$y^* < \theta \mid C_x \in \{C_{1,\ldots,S}\} \tag{6}$$

Thus, a miss occurs for a target input (i.e., an input belonging to one of the target classes $C_1,\ldots,C_S$) when the maximum score y^* is below threshold. However, for target inputs and $k<S$, the multitarget detector is subject to another type of error that occurs when y^* is above threshold but C_x, the class of the input x, is not one of the top-k hypotheses:

$$y^* > \theta, C_x \notin \{h_k^*\} \mid C_x \in \{C_{1,\ldots,S}\} \tag{7}$$

This condition represents a confusion error in that one of the detector scores exceeds threshold but the class associated with the correct score is not one of the top-k hypotheses. (When $k=S$, the input class is always among the top-k hypotheses and there are no confusions.) In our approach we count both types of errors (from Eqs. (6) and (7)) as misses so that

$$Miss':\ [(y^* < \theta) \cup (y^* > \theta, C_x \notin \{h_k^*\}] \mid C_x \in \{C_{1,\ldots,S}\}$$

(8)

We note that in our formulation the top-k list is produced whenever $y^*>\theta$, regardless of whether the remaining k-1 scores are above the threshold. This approach differs from that developed by others [1].

A taxonomy of possible multitarget detector decisions is shown in Figure 3. The errors labeled "False Alarm" and "False Reject" are analogous to the false alarm and miss errors in the single detector system. The second type of miss error ("Confusion"), shown in the shaded portion of the figure, occurs when the input is of a target class, the maximum score is above threshold, but the class of the input is not one of the top-k hypotheses. Note that a correct accept decision occurs whenever the correct class C_x is among the top-k hypotheses regardless of whether the associated score is highest.

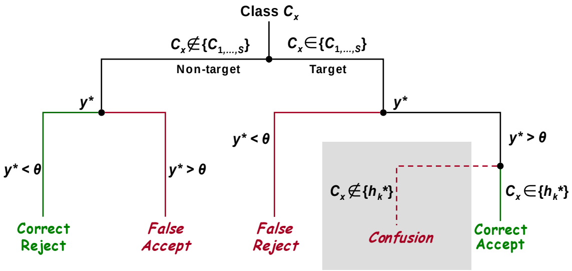

Figure 3: Taxonomy of multitarget detector decisions.

2.4. Performance

Detection performance is typically reported by plotting the miss rate as a function of the false alarm rate at all possible score thresholds. False alarm rates are computed from the proportion of non-target trials (input class C_x does not equal detector class C) whose scores exceed threshold and miss rates are computed from the proportion of target trials (input class C_x equals detector class C) whose scores fall below threshold. Reporting performance for multitarget detectors requires two extensions. First, a target trial now refers to the case where the class of the test message is one of the S stack detector classes $C_{1,\ldots,}C_S$. Second, the detection error tradeoff becomes a function of k and a set of S DET plots for k=1,...,S is required to characterize detection performance fully. Since performance is bounded by the top-1 and top-S DETs, we will tend to focus our experiments on the top-1 and top-S conditions.

Speaker and language multitarget recognition systems may also be characterized by their closed set identification rates, which measure the proportion of test messages whose class is represented in the top-k hypotheses. Since no imposter messages are tested for identification, no false alarm rate is reported.

3. Predicting multitarget detector performance

In this section we derive equations for the stack detector's probability of miss and false alarm errors using the miss and false alarm probabilities of the single detector. We first look at the bounding cases of top-*S* and top-1 stack detectors and then generalize to the top-k stack detector. For simplicity of the final equations, we will assume in all cases that the detectors operate independently of each other, that they all have the same miss and false alarm probability characteristics $P_{miss}(\theta)$ and $P_{fa}(\theta)$ as defined in equations (3) and (4), and that the prior probabilities of the target classes are equal. Under these assumptions, $P_{miss}(\theta)$ and $P_{fa}(\theta)$ are referred to as the miss and false alarm probabilities of the *prototype* single target detector. In the Appendix, we derive the general expressions for the top-*S* stack detector probability of miss and probability of false alarm.

3.1. Top-*S* stack detector

For a stack detector of size *S*, a false alarm occurs if one or more of the individual detectors false alarm. Since the false alarm probabilities of the individual detectors are assumed to be equal, the probability of this event can be expressed as the complement of the probability that none of the detectors false alarm

$$P'_{fa}(\theta) = 1 - (1 - P_{fa}(\theta))^S \qquad (9)$$

where $P'_{fa}(\theta)$ is the notation used for the probability of false alarm of a stack detector. Note that $P'_{fa}(\theta)$ is independent of the value of *k* in the top-*k* decision rule.

For the top-*S* case, a miss occurs only under the conditions of Eq. (6), that is, when the detector that corresponds to the input class C_x misses and none of the other *S*-1 detectors false alarm. Since the miss and false alarm probabilities of the individual detectors are assumed to be equal, the probability of this event can be expressed as

$$P'_{miss}(\theta, S) = P_{miss}(\theta) * (1 - P_{fa}(\theta))^{S-1} \qquad (10)$$

where $P'_{miss}(\theta, S)$ is the notation used for the probability of miss of the top-*S* stack detector. There are no confusions in a top-*S* stack detector because the correct class will always be among the top-*S* hypotheses.

Figure 4 shows a plot of Eqs. (9) and (10) for stack sizes *S*=1,…,10 for a stack detector composed of prototypes operating at $P_{miss}(\theta) = P_{fa}(\theta) = 0.1$. We see that the false alarm rate increases dramatically with increasing stack size. As has been noted by others [2], this behavior implies that individual detectors must operate at very low false alarm rates for a stack detector with a large stack size to have a reasonable false alarm rate. Also notable is the decrease in miss rate with increasing stack size, a consequence of using a top-*S* rule where a detect by any detector, not just the one corresponding to the input test class, counts as a correct detect. While this is a reasonable rule for counting misses for a small stack size (e.g., a *group detector* where it is of interest if any member of a small group is present), the rule becomes less useful as the stack size increases.

3.2. Top-1 stack detector

For the top-1 stack detector, the false alarm probability is the again given by Eq. (9) since the top-*k* decision rule has no impact on false accept errors. From Eq. (8) we see that a top-1 stack detector miss can occur for target inputs when either the maximum detector score *y** is below the detection threshold or when *y** is above the threshold but the top-1 detector hypothesis h^* does not match the input class C_x. These two conditions are defined in Eqs. (6) and (7). The probability of the first condition is the same as is given in Eq. (10). Assuming independence, the probability of the second condition consists of the complement of Eq. (10) multiplied by the probability that the score of the detector of class C_x is not the maximum score. Following the method proposed in [1, Appendix 5], we use a moment model to compute this probability.

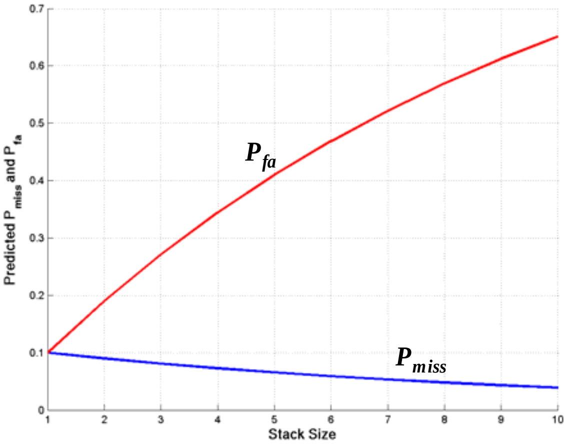

Figure 4: Plot of top-S stack detector miss and false alarm probabilities for various stack sizes, with all individual detectors operating at $P_{miss}(\theta) = P_{fa}(\theta) = 0.1$

Let $p_{tar}(\tau)$ be the probability density function (pdf) of scores *y* from a prototype detector when the input class C_x matches that of the detector, and let $p_{non}(\tau)$ be the pdf of prototype detector scores *y* when the input class C_x does not match that of the detector

$$p_{tar}(\tau) = p_y(\tau \mid C_x = C) \qquad (11)$$

$$p_{non}(\tau) = p_y(\tau \mid C_x \neq C) \qquad (12)$$

Consider a stack detector of size *S*=2 composed of two prototype detectors. If the score y_1 of the target detector (class C_1) is τ, then the probability that the score y_2 of the non-target detector is less than τ is

$$\Pr[y_1 > y_2 \mid y_1 = \tau, C_x = C_1] = \int_{-\infty}^{\tau} p_{non}(\alpha) d\alpha = 1 - P_{fa}(\tau) \qquad (13)$$

Integrating this expression over the target score pdf gives us the probability that the target detector score will be the maximum for any τ

$$\Pr[y_1 > y_2 \mid C_x = C_1] = \int_{-\infty}^{\infty} \left(1 - P_{fa}(\tau)\right) p_{tar}(\tau) d\tau \qquad (14)$$

With S detectors (one target detector and S-1 non-target detectors) in the stack, Eq. (14) generalizes to

$$\Pr[y_i = \max(y_1,\ldots,y_S) \mid C_x = C_i] = \int_{-\infty}^{\infty} \left(1-P_{fa}(\tau)\right)^{S-1} p_{tar}(\tau)d\tau \qquad (15)$$

The probability that the target detector score is *not* the maximum score is then the complement of equation (15).

Putting all of the above together, we can write the probability of miss for a top-1 stack detector as

$$P'_{miss}(\theta,1) = P_{miss}(\theta)*(1-P_{fa}(\theta))^{S-1} + \left(1-P_{miss}(\theta)*(1-P_{fa}(\theta))^{S-1}\right)* \left(1-\int_{-\infty}^{\infty}\left(1-P_{fa}(\tau)\right)^{S-1} p_{tar}(\tau)d\tau\right) \qquad (16)$$

In Figure 5 we plot the probability of miss as a function of threshold θ for a stack detector of size 10 for top-S and top-1 conditions using Eqs. (10) and (16). The probability of miss of the single detector prototype is shown for reference. The pdfs were Gaussian fits to speaker detection score data and numerical integration was used to evaluate Eq. (16). From the equations and the plot we see that the top-1 miss rate matches the top-S miss rate for high values of θ but then smoothly asymptotes to the closed-set identification error rate as θ gets smaller. This is expected behavior since the tests that give rise to confusions (score of target detector not the maximum score) are misses that cannot be recovered by a change of threshold. We also see that the miss rate of the top-1 stack detector is generally higher over all thresholds since a correct detect is now more difficult to achieve.

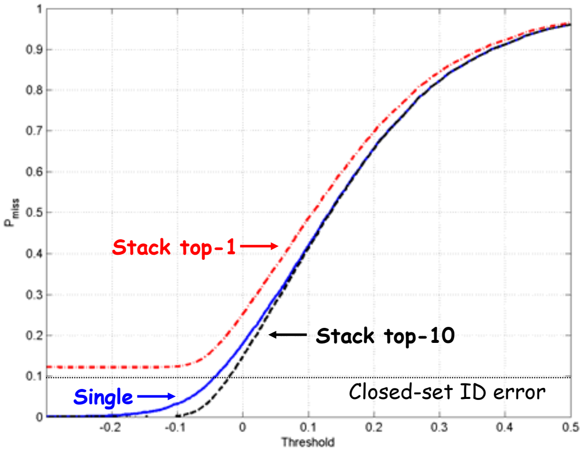

Figure 5: Plot of P_{miss} as a function of threshold for single detector (solid), stack detector top-S (dashed) and stack detector top-1 (dashed-dot). Stack size is 10.

3.3. Top-*k* stack detector

To generalize to the top-k case, we again follow the moment model used in [1]. In this case, if the maximum score is above threshold, the detector corresponding to the test class must be among the top-k hypotheses for a correct detection. The probability that the target detector score is among the top-k scores can be computed by enumerating all cases in which this occurs and using the appropriate binomial coefficients

$$\Pr[y_i \in \max_k(y_1,\ldots,y_S) \mid C_x = C_i] = \sum_{j=1}^{k}\int_{-\infty}^{\infty}\binom{S-1}{j-1}\left(P_{fa}(\tau)\right)^{j-1}\left(1-P_{fa}(\tau)\right)^{S-j} p_{tar}(\tau)d\tau \qquad (17)$$

The probability that the target score is not among the top-k scores is the complement of Eq. (17). The probability of miss for the general top-k stack detector is then

$$P'_{miss}(\theta,k) = P_{miss}(\theta)*(1-P_{fa}(\theta))^{S-1} + \left(1-P_{miss}(\theta)*(1-P_{fa}(\theta))^{S-1}\right)* \left(1-\sum_{j=1}^{k}\int_{-\infty}^{\infty}\binom{S-1}{j-1}\left(P_{fa}(\tau)\right)^{j-1}\left(1-P_{fa}(\tau)\right)^{S-j} p_{tar}(\tau)d\tau\right) \qquad (18)$$

In the experiments described in the next section, we will focus on top-1 and top-S stack detectors since they provide the bounding cases of interest.

4. Experiments

In this section we describe the experiments that were designed to compare stack detector performance results predicted using the equations of Section 3 with actual performance. Experiments were performed on both speaker and language corpora.

Multitarget language detection was performed using a GMM based language recognition system identical to the one described in [5] but without gender dependent models or a backend. Each detector score was computed as the ratio of the detector GMM output to the average of the remaining 11 GMM outputs (scores were allowed to interact for language detection). Recognition scores were generated using the 1996 12-language NIST development set (“lid96d1”) [6]. This corpus consists of approximately 1200 30s messages, with roughly 160 messages each for English, Mandarin, and Spanish and 80 messages for each of the other nine languages. Miss and false alarm statistics were obtained by pooling the scores across the 12 languages and calculating miss and false alarm rates using all scores as thresholds. These rates are used to characterize the prototype language detector. The 12-language pooled EER on the test corpus is 12.2%.

Multitarget speaker detection was performed using a GMM-UBM speaker detection system described more fully in [7]. In this system feature mapping [8] was employed and a UBM was built using an agglomeration of cellular and landline telephone data from Switchboard-II phase 1 and phase 4 corpora as well as the OGI Cellular database. To produce a large number of speakers and tests, the union of speakers and tests from the 1998, 1999, and 2002 NIST speaker recognition evaluations was used for experiments (where the data comes from Switchboard-II phase 2, phase 3 and phase 5 cellular). The aggregate set consisted of 1369 speaker models (619 males and 750 females) and 12648 test utterances (5780 males and 6863 females). All male speaker models were scored against all male test utterances to give a 5780 x 619 score matrix from which to operate (containing 5433 target trials). Similarly we had a 6863 x 750 score matrix for the female speakers (containing 6361 target trials). Log-likelihood ratio scores were computed between the target model and the UBM, with no interaction among the speaker models for score computations. The pooled (prototype) EER is 8.4% for the male speakers and 8.8% for the female speakers. Empirical performance for an S speaker multitarget detector was computed using a Monte Carlo simulation wherein 100 sets of S speakers were randomly selected and counts of misses and false alarms were accumulated and used to compute performance. Experiment results are shown using the male speaker models and test utterances only.

4.1. Prediction of top-*S* miss and false alarm probabilities

Our first goal was to establish the validity of the predictions of top-S multitarget detector miss and false alarm probabilities (Eqs. (9) and (10)). Using the 12 individual language detectors we computed the actual miss and false alarm rate curves for all 220 possible stacks of 3 languages. Next, we used Eqs. (9) and (10) to compute the predicted stack detector miss and false alarm probabilities. Results are shown in Figure 6, where black lines are the predicted top-S stack detector miss and false alarm probabilities and red points are the 220 measured error rate curves. The measured prototype rates are shown in blue for reference. Correspondence between the predicted and mean measured rates appears to be quite good, but note that there is a significant amount of variability in the individual measured results that is not addressed by the predictions.

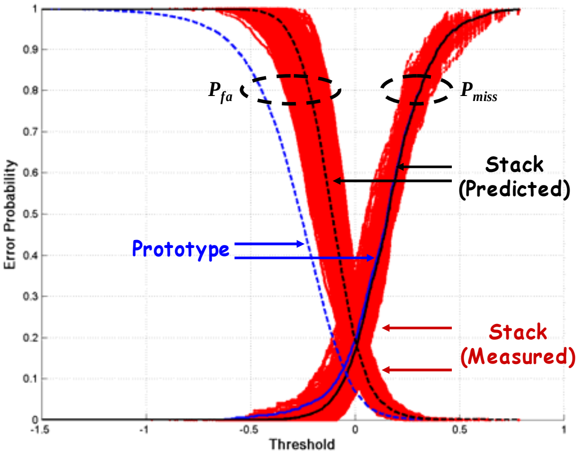

Figure 6: Measured and predicted top-S miss and false alarm probabilities for a 3-detector language stack.

4.2. Prediction of top-*S* multitarget DET plots

Figure 7 compares measured and predicted DET plots for multitarget language recognition with stack size S=3. The DET plot for the multitarget detector was obtained by pooling the decisions of all 220 possible 3-language runs. Predicted results are close to measured results at low false alarm rates (high threshold) but the curves diverge somewhat as the false alarm rate increases (threshold decreases). At these higher rates, measured detection performance is slightly better than predicted. The top-S EER for the average 3-language stack detector is 17.7% while that of the single detector prototype is 12.2%.

Figure 8 illustrates the manner in which points on the prototype DET map to the multitarget DET and vice versa. A stack detector built of three prototypes, each of which operates at an EER of 12.2%, will have an operating point at P'_{fa}=32.3% and P'_{miss}=9.4%. Similarly, a 3-language stack of detectors with a desired EER of 17.7% must be composed of prototypes operating at P_{fa}=6.3% and P_{miss}=20.2%. These correspondences illustrate the point made in Section 3 and Figure 4 that the false alarm rate of a multitarget detector will be considerably greater than that of its component detectors.

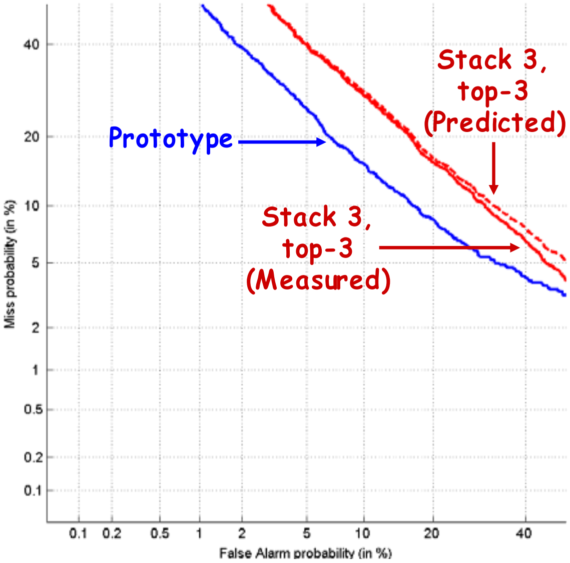

Figure 7: DET plots for multitarget language recognition with stack size S=3. The dashed red line shows 3-language multitarget detector performance predicted from the single detector prototype (solid blue line). The solid red line is plotted from the multitarget detector measured data.

Figure 9 and Figure 10 show predicted and measured top-S language and speaker recognition performance for varying stack sizes. The speaker results in Figure 9 are shown for stack sizes S=10 and S=100 and were obtained by pooling the results of 100 Monte Carlo runs, as described above. The multitarget language results in Figure 10 are shown for S=2 and S=6 and were obtained in a manner similar to those of Figure 7. For both speaker and language we find that prediction of multitarget detection performance is quite accurate and that increasing the stack size leads to an increase in false alarm rates, as expected.

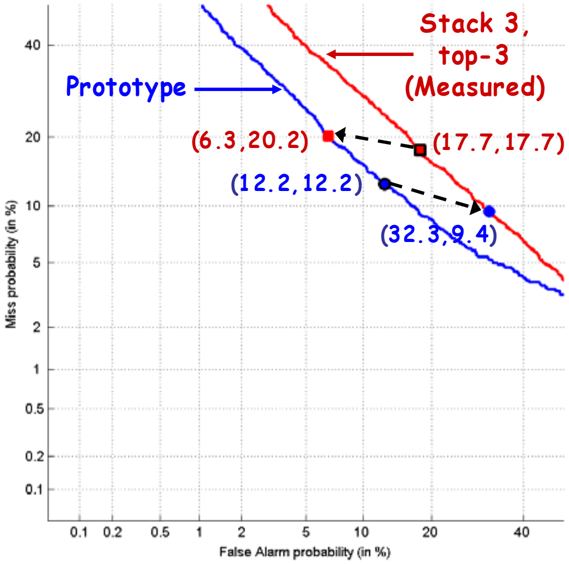

Figure 8: Correspondence between single detector (blue) and multitarget detector (red) operating points.

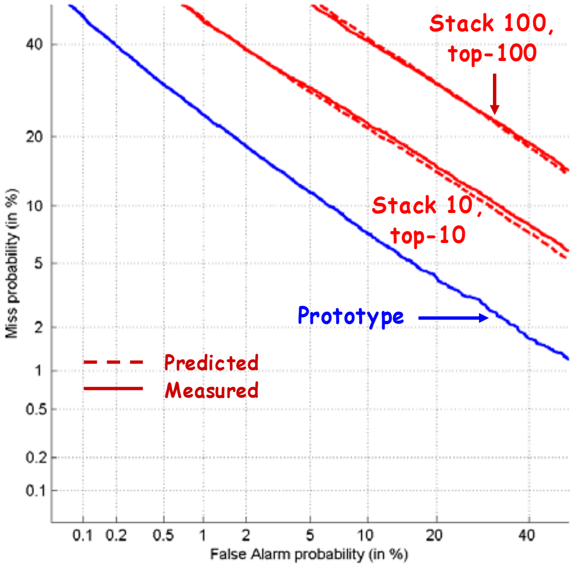

Figure 9: Comparison of predicted DETs (dashed lines) and actual DETs (solid lines) for top-S multitarget speaker detection for stack sizes 10 and 100. DET of single target prototype is the lower blue line.

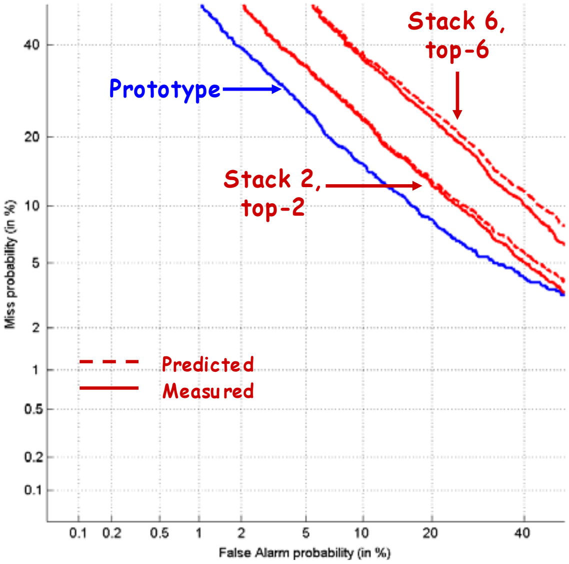

Figure 10: Comparison of predicted DETs (dashed lines) and actual DETs (solid lines) for top-S multitarget language detection for stack sizes 2 and 6. DET of single target prototype is the lower blue line.

4.3. Top-1 multitarget detection

Figure 11 and Figure 12 show a comparison of top-*S* and top-1 DET plots for multitarget speaker and language detection. As described in Section 3.2, top-1 systems classify as misses those target trials for which the maximum score y^* exceeds threshold but for which the correct class is not h^*. These errors, which represent confusions by the multitarget detector, cannot be recovered by lowering the threshold. Consequently, the miss rate is subject to a lower bound given by the top-1 confusion rate (the closed-set identification error rate). For a given stack size, we expect the top-*S* and top-1 DET plots to diverge at increasing false alarm rates (decreasing θ), with the top-1 DET approaching an asymptote given by the top-1 confusion rate. This prediction is confirmed in Figure 11 which shows the top-*S* and top-1 measured DETs for a size 10 stack of speaker detectors. The top-10 DET is identical to that shown in Figure 9 for *S*=10 whereas the top-1 DET approaches the confusion rate asymptote (measured at 9.6% for this experiment). The predicted top-1 DET for this experiment is shown by the dashed red line in Figure 11. Figure 12 shows the top-*S* and top-1 measured DETs for a size 3 stack of language detectors. The top-3 DET is identical to that shown in Figure 7 and the top-1 DET approaches the confusion rate asymptote (measured at 11.1% for this experiment) at high false alarm rates. The dashed red line shows the top-1 performance of the stack detector as predicted by Eqs. (9) and (16). For both speaker and language recognition, predicted top-1 multitarget detection performance underestimates actual performance (i.e., actual performance is better than predicted) and does not match measured results as accurately as top-*S* performance did.

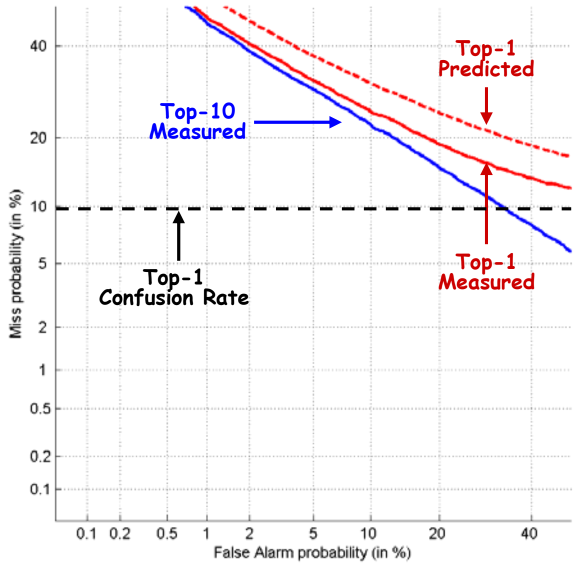

Figure 11: The measured top-S (blue), measured top-1 (red), and predicted top-1 (dashed red) DET plots for multitarget speaker recognition with S=10. The dashed horizontal line shows the top-1 closed-set confusion rate (9.6%).

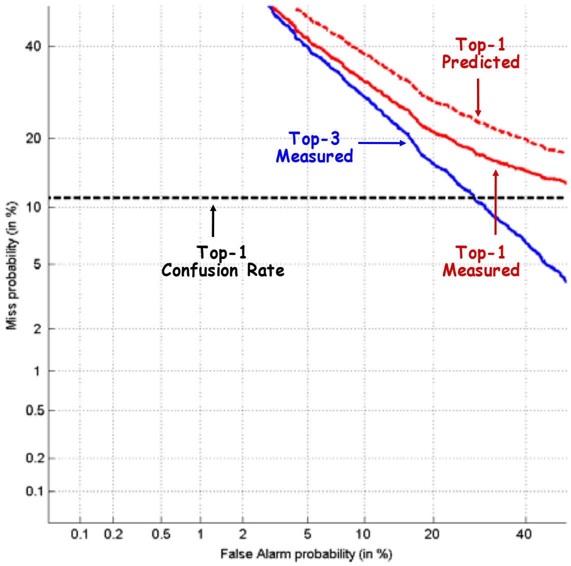

Figure 12: The measured top-S (blue), measured top-1 (red), and predicted top-1 (dashed red) DET plots for multitarget language recognition with S=3. The dashed horizontal line shows the top-1 closed-set confusion rate (11.1%).

4.4. Prediction of closed-set identification error

In Eq. (15) of Section 3.2 we derived an expression for the probability that the target detector produces the maximum score for an input whose class is one of the target classes. The complement of this equation represents the closed-set identification error rate (or confusion rate) and is given by

$$P'_{conf}(S) = 1 - \int_{-\infty}^{\infty} \left(1 - P_{fa}(\tau)\right)^{S-1} p_{tar}(\tau) d\tau \qquad (19)$$

Theoretically, then, it is possible to predict the confusion rate of a stack detector of any size given the probability density functions $p_{tar}(\tau)$ and $p_{non}(\tau)$ of the prototype detector.

Figure 13 and Figure 14 show plots of predicted vs. measured confusion rates for both language and speaker recognizers. Accurate predictions are obtained for the language recognizer over its relatively limited range of stack sizes. For the speaker case, where the stack confusion rates were measured for stack sizes up to *S*=618, the predictions significantly overestimate the confusion rates. It is not clear at this time why the confusion rate predictions are inaccurate, although we note that similar inaccuracies were observed in face recognition experiments [1]. The assumption of independence in the detectors and the approximations used to compute the integrals are both possible sources of errors.

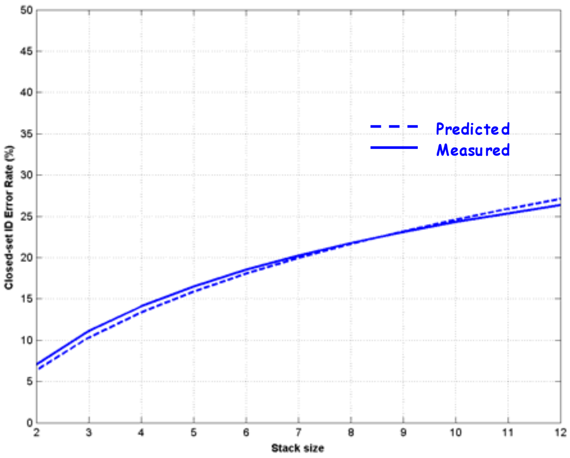

Figure 13: Predicted (dashed line) and measured (solid line) closed-set identification error rates for multitarget language recognition for stack sizes S=2,…,12.

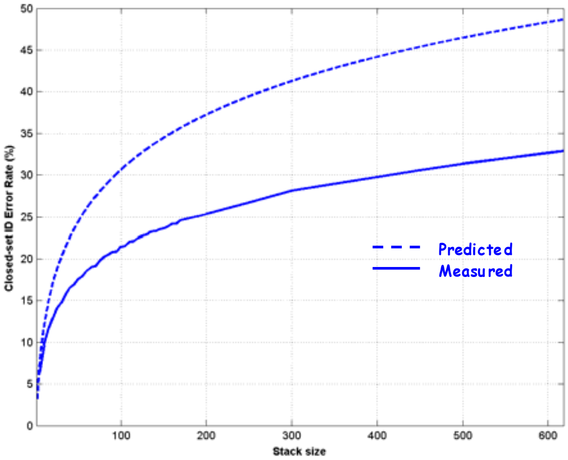

Figure 14: Predicted (dashed line) and measured (solid line) closed-set identification error rates for multitarget speaker recognition for stack sizes S=2,…,618.

5. Conclusions

In this paper we presented a framework for the study of multitarget detectors for speaker and language recognition applications. We showed that analysis of the behavior of a multitarget detector requires a generalization of our understanding of the single target detector with regard to the decisions that are produced and to the types of errors that are generated. Specifically, multitarget detection was regarded as a two stage process comprising detection (comparison of the maximum score to a threshold) and identification (selection of a set of hypotheses). An analysis of the possible outcomes of multitarget detection indicated that a new type of error, referred to as a confusion, must be included in the description of the detector's performance. Our approach was to include this error in the calculation of the detector miss rate.

Our analysis allowed us to derive formulas to predict the miss and false alarm rates of a multitarget detector from the error characteristics of the individual detectors in the stack. As has been shown by others, the analysis indicates that false alarm performance degrades rapidly as the size of the stack increases. Using data from both speaker and language corpora, we showed that the predicted top-*S* results match the measured results quite well over a range of stack sizes but that prediction of top-1 and closed-set identification results was somewhat less accurate, perhaps due to erroneous assumptions regarding the independence of the detectors or to inaccurate numerical integration techniques. These deviations from expectations will be the subject of further investigation. For top-*S* detection applications, we may conclude that the performance of a multitarget detector is predictable from the performance of its individual detectors and that research focused on improving single target detector performance can be mapped directly to the top-*S* multitarget detection task.

6. Appendix

6.1. Top-S stack detector: General case

In this section we derive the general expressions for the top-*S* stack detector probability of miss and false alarm. Assume a stack of *S* independent detectors in which the probability of miss and probability of false alarm of the *i*th detector are given by $P^{(i)}_{miss}(\theta)$ and $P^{(i)}_{fa}(\theta)$, and that the input classes occur with prior probability p_i. Since a false alarm occurs if any single detector false alarms, the probability of the stack detector producing a false alarm is the complement of the probability that no single detector false alarms:

$$P'_{fa}(\theta) = 1 - \prod_{i=1}^{S} [1 - P^{(i)}_{fa}(\theta)] \qquad (20)$$

If all detectors have the same false alarm probability, Eq. (20) reduces to Eq. (9).

For a top-*S* stack detector we use Eq. (6) to write the probability of miss as

$$\begin{aligned} P'_{miss}(\theta, S) &= \Pr[y^* < \theta \mid C_x \in C_{1,\ldots,S}] \\ &= \Pr[y^* < \theta, C_x \in C_{1,\ldots,S}] / \Pr[C_x \in C_{1,\ldots,S}] \\ &= \Pr[y^* < \theta, C_x \in C_{1,\ldots,S}] / (\sum_{i=1}^{S} p_i) \end{aligned}$$

Assuming independence, this becomes

$$\begin{aligned} P'_{miss}(\theta, S) &= (1/\sum_{i=1}^{S} p_i) \sum_{i=1}^{S} \Pr[y^* < \theta, C_x = C_i] \\ &= (1/\sum_{i=1}^{S} p_i) \sum_{i=1}^{S} \Pr[y^* < \theta \mid C_x = C_i] \Pr[C_x = C_i] \\ &= (1/\sum_{i=1}^{S} p_i) \sum_{i=1}^{S} p_i \Pr[y^* < \theta \mid C_x = C_i] \end{aligned}$$

(21)

Since the input class C_x is of the target class C_i, Eq. (21) can be expressed in terms of the individual detector probabilities and class priors as

$$P'_{miss}(\theta, S) = (1/\sum_{i=1}^{S} p_i) \sum_{i=1}^{S} p_i P^{(i)}_{miss}(\theta) \prod_{j \neq i} [1 - P^{(j)}_{fa}(\theta)]$$

(22)

When $P^{(i)}_{miss}(\theta) = P_{miss}(\theta)$ and $P^{(i)}_{fa}(\theta) = P_{fa}(\theta)$, Eq. (22) becomes

$$P'_{miss}(\theta, S) = P_{miss}(\theta) * (1 - P_{fa}(\theta))^{S-1}$$

as in Eq. (10).

7. References

[1] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and J.M. Bone, "FRVT 2002: Evaluation Report." March 2003, http://frvt.org/FRVT2002/documents.htm

[2] J. Daugman, "Biometric Decision Landscapes.' *Technical Report No. TR482, University of Cambridge Computer Laboratory*. http://www.cl.cam.ac.uk/users/jgd1000

[3] A. Martin et al., "The DET Curve in Assessment of Detection Task Performance." *Proc. Eurospeech '97*, pp. 1895-1898.

[4] P. Grother, R. Micheals, and P.J. Phillips, "Face Recognition Vendor Test 2002 Performance Metrics."

http://www.frvt.org/DLs/Avbpa_2003_evaluation_metrics.pdf

[5] E. Singer et al., "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification." *Proc. Eurospeech 2003*, pp. 1345-1348.

[6] http://www.nist.gov/speech/tests/lang/index.htm

[7] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models." *Digital Signal Processing*, vol. 10, January 2000, pp. 19-41.

[8] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping." *Proc ICASSP 2003*, pp. 53-56.